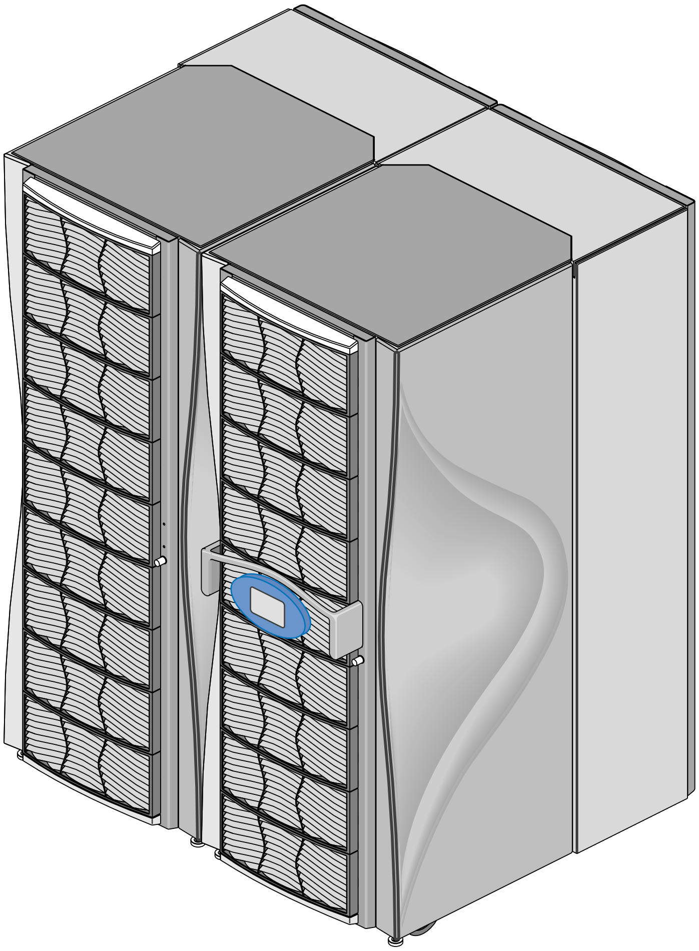
SGI™ 3000 Family

Reference Guide



SGI™ 3000 Family Reference Guide

Introduced in July 2000, the SGI 3000 family of high-performance servers and visualization systems is the latest addition to the line of distributed shared-memory systems from SGI. Known as the SGI™ Origin™ 3000 line of servers and the SGI™ Onyx(R) 3000 line of visualization systems, all members of the SGI 3000 family are designed with the SGI™ NUMA 3 architecture, the third generation in a line of Non-Uniform Memory Access (NUMA) architectures from SGI.

Starting from a minimum configuration of two 64-bit MIPS(R) RISC microprocessors, the NUMA 3 architecture of the SGI 3000 family scales to support up to 512 processors in a single memory image, single operating system kernel, cache-coherent multiprocessor system. In addition to being a highly scalable server platform, the SGI InfiniteReality3™ graphics subsystem is tightly integrated into the architecture to provide customers with a highly scalable visualization system.

There are a total of six models in the SGI 3000 family--three configurations are specified as servers and another three as visualization systems. Each model is defined by the maximum number of processors supported in the system.

- SGI 3200 series--supports a maximum of 8 processors
    - SGI™ Origin™ 3200
    - SGI™ Onyx(R) 3200
- SGI 3400 series--supports a maximum of 32 processors
    - SGI™ Origin™ 3400
    - SGI Onyx 3400
- SGI 3800 series--supports a maximum of 512 processors
    - SGI™ Origin™ 3800
    - SGI™ Onyx(R) 3800

SGI 3000 Family Product Overview

With the announcement of the SGI 3000 family, SGI introduces NUMAflex^{TM} as the desired methodology for creating high-performance systems to meet the vast information demands of the future. NUMAflex offers the ability to configure a modular and scalable system to meet the demands of your application environment. The concept of NUMAflex delivers a proven SGI NUMA architecture and a bevy of benefits for the customer, namely maximum flexibility, resilience, and investment protection:

- Delivers precise mix of capabilities, saving money and space
- Virtually unlimited expansion takes the guesswork out of planning
- Reconfigures to meet changing requirements without penalty
- Breakthrough in availability, service, administration, and logistics
- Incorporate new technologies as they're introduced (e.g., IA-64, I/O)
- Independent component upgrades for maximum investment protection
- Isolate components as necessary for easy serviceability
- Run your machine in single system image (shared-memory) or cluster mode
- Manage your large SSI system, partitioned system, or cluster from a single point of administration using a management platform from SGI
- Use NUMAlink$_{TM}$ as a high-speed interconnect to run shared memory, message passing, or OpenMP^{TM} jobs, and as a capability or throughput machine simply through software changes--no cabling changes are required
- Customers can buy, deploy, and redeploy building blocks to build the correct system and exact configuration for the job they wish to do
- Modular and reconfigurable system allows easy upgrades as needed when new technology is introduced
- NUMA 3 modularity allows customers to remove and service failed components while the rest of the system continues to produce results
- Software enhancements offer increased control of partitioning and automatic restart of system

NUMA 3 Architecture

The system architecture for the SGI 3000 family is a third-generation NUMA architecture from SGI known as NUMA 3. In the SGI NUMA 3 architecture, all processors and memory are tied together into a single logical system through the use of special crossbar switches developed by SGI. This combination of processors, memory, and crossbar switches constitute the interconnect fabric called NUMAlink.

One component of the SGI NUMA 3 architecture is the Bedrock chip. This ASIC is an 8-input by 6-output crossbar that acts as the memory controller between processors and memory in the system, both local and remote. The Bedrock chip has a total aggregate peak bandwidth of 3.2GB/second. The chip also has a channel that connects processors to system I/O, which allows every processor in a system direct access to every I/O slot in the system.

Another component of the SGI NUMA 3 architecture is the router chip. This 8-port crossbar ASIC is found in a router node and, through the use of highly specialized cables, provides the high-bandwidth, extremely low latency interconnect known as NUMAlink. The router node channels information between all the compute nodes in the system, connecting all Bedrock switches together to create a single contiguous memory in the system of up to 1 terabyte.

These crossbar switches, together with the NUMAlink interconnect fabric, deliver an extremely low-latency and high-bandwidth architecture, not just another cluster solution tightly integrated in a single chassis and passed off as an SMP system. From the smallest 4-processor system to the largest 512-processor supercomputer, the memory latency ratio between remote and local memory in the SGI NUMA 3 architecture is only 2:1 and still less than 600 nanoseconds round-trip in the largest configuration.Not all NUMA implementations are considered equal. Other NUMA architectures available in the industry exhibit remote to local memory latencies of 3:1 or worse, proving how difficult it is to design an effective NUMA implementation. While other system vendors introduce NUMA architectures for the first time, SGI is delivering its third generation and has nearly 4 years of experience delivering advanced NUMA solutions.

The SGI NUMA 3 architecture provides the flexibility necessary to build a series of highly scalable servers and high-end visualization systems. The SGI 3000 family consists of modular systems built by using SGI NUMA building blocks installed in industry-standard 19-inch racks. Two rack sizes are offered as a starting point from which the customer can configure a system. For smaller configurations or situations where space is limited, a 17U deskside rack is recommended. For larger systems where scalability is required, a

39U full-height rack can be used to build systems capable of supporting thousands of processors. A 16-processor block diagram of the SGI NUMA 3 architecture is illustrated in Figure 1.

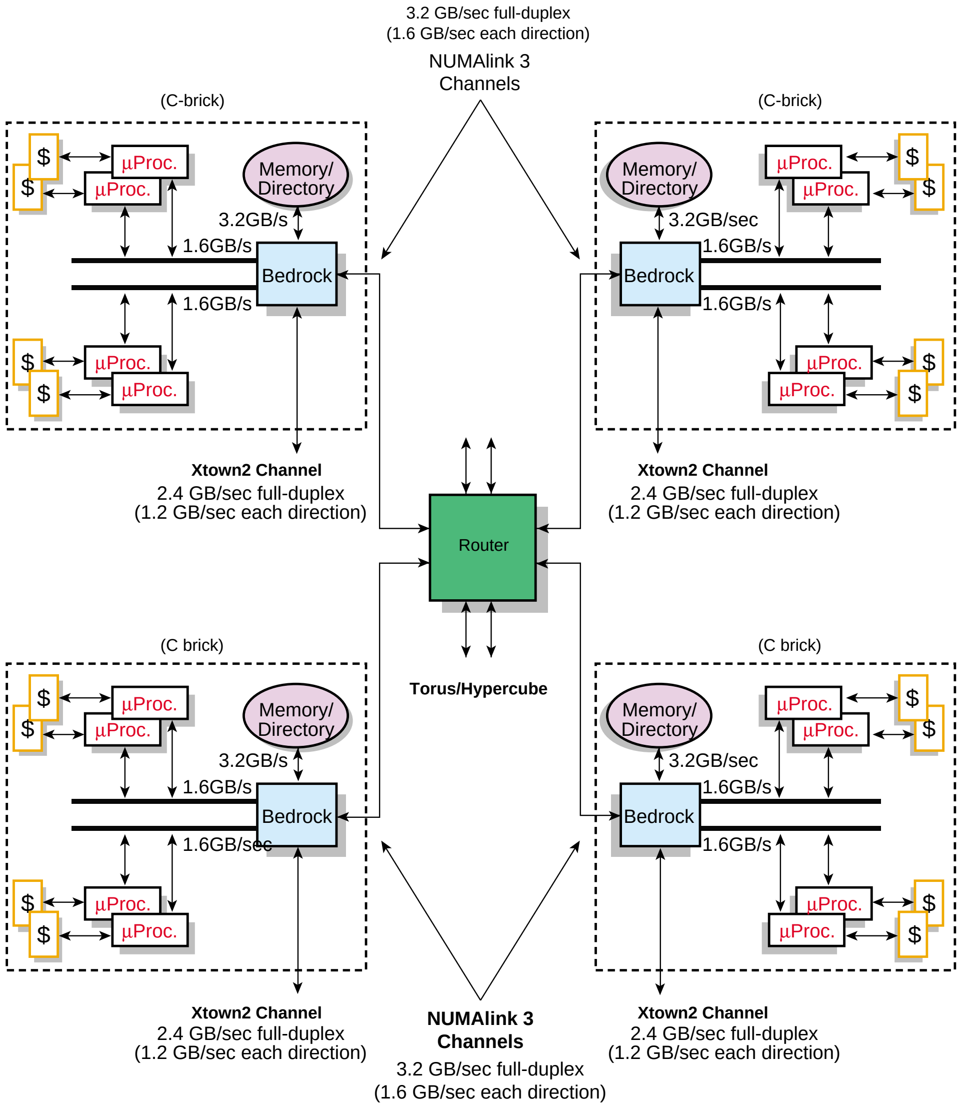


Figure 1. SGI 3000 Family 16-Processor Configuration

SGI NUMA 3 Building Blocks-Bricks

All system configurations are generated by combining SGI NUMA building blocks called “bricks.” Each brick provides a specific function and can be added independently of other functional bricks in the system to achieve a design specifically addressing your application requirements. As bricks are added to a system, the bandwidth and performance scale in a manner that is almost linear.

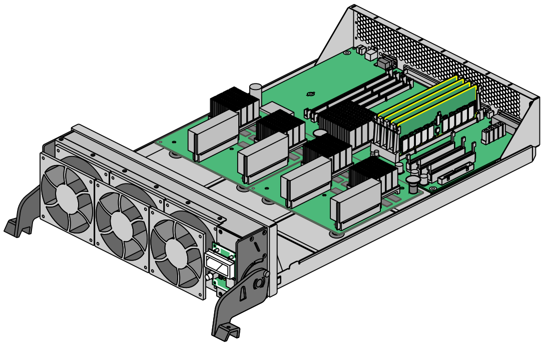
The following is a list of bricks used to build the SGI 3000 family of systems:

• C-brick	CPU module (processors and memory)
• R-brick	Router interconnect
• I-brick	Base I/O module
• P-brick	PCI expansion
• X-brick	XIO expansion
• G-brick	Graphics expansion
• D-brick	JBOD disk storage
• Power bay	N+1 redundant power

C-brick

Features:

- Two or four processors
- Four memory banks
- Bedrock memory controller with 6.4GB/second aggregate bandwidth
- 512MB to 8GB memory
- 512MB, 1GB, 2GB memory options



The C-brick is a 3U rack-mountable enclosure that houses processors and memory in the system. Scaling the compute performance of the system is as simple as adding more C-bricks. Each C-brick contains two or four MIPS 64-bit processors, and CPU to memory bandwidth is controlled by the Bedrock switch. Two MIPS processors share a single 1.6GB/second channel to the Bedrock, and the Bedrock has two such channels, so an aggregate of up to 3.2GB/second of CPU to memory bandwidth is available for local memory addressing per C-brick throughout the system. Refer to Figure 2, which illustrates two C-bricks in an 8-processor configuration.

Each C-brick includes one network interface (NI) and one I/O interface (II). These interfaces are used to attach NUMAlink interconnect cables, which in turn integrate each C-brick into the interconnect fabric. The NI channel connects the C-brick to a router in the system, and the II channel connects to a single I/O brick.

C-bricks within SGI NUMA 3 will support a mixture of processor speeds in the system, where each processor runs at its own native speed. This feature allows you to upgrade as necessary by adding newer, faster processors without replacing all existing processors and benefit from increased performance with the latest technology.

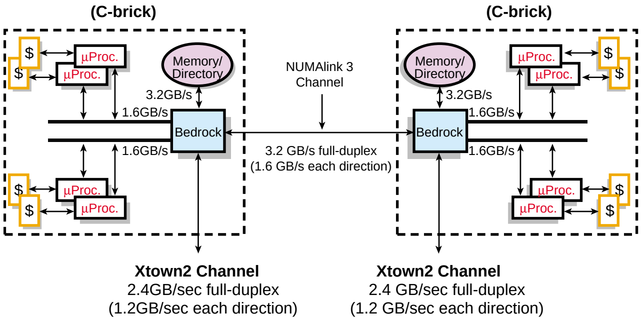
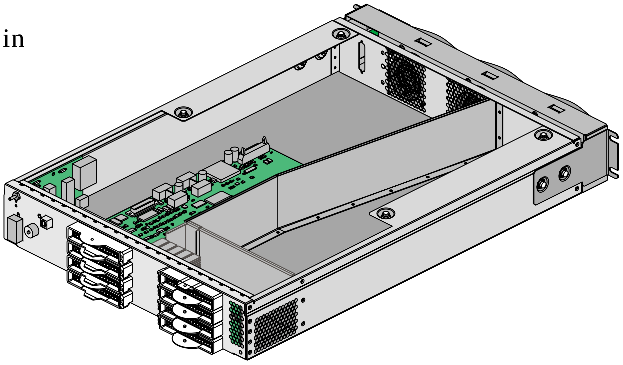


Figure 2. SGI 3000 series 8-processor configuration

R-brick



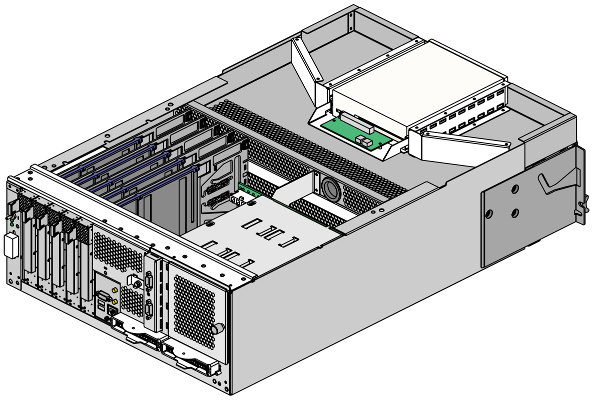
The R-brick is the foundation of the SGI NUMA 3 architecture used in the SGI 3000 family. The R-brick is a 2U rack-mount enclosure that acts as a central hub connecting all C-bricks in the SGI NUMA 3 interconnect fabric. Each R-brick provides eight NUMAlink channels, one for every port of the router crossbar. Four channels are used to connect to four C-bricks, so up to 16 processors can be interconnected to a single R-brick. The remaining four channels are responsible for communication to other Router nodes in the system. Each router node can communicate with similar router nodes, each supporting 16 processors, up to a maximum configuration of 128 compute nodes or 512 processors.

There are three versions of the R-brick:

- 6-port router-standard in SGI Origin 3400 and SGI Onyx 3400, can be configured to support systems up to 32 processors
- 8-port Router-standard in SGI Origin 3800 and SGI Onyx 3800, can be configured to support up to 128 processors
- Metarouter-for the largest configurations, supporting up to 512 processors in single shared-memory system (required on systems over 128 processors)

The R-brick, like all other bricks in the SGI NUMA 3 architecture, can be independently upgraded at any point in the future. This is critical for your system to support additional bandwidth and functionality required for advanced bandwidth needs of the future.

I-brick

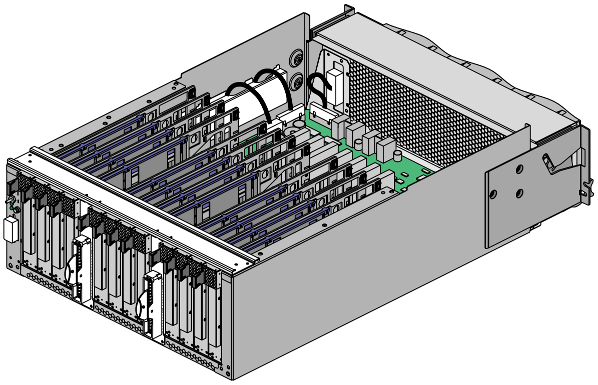


The I-brick contains the base I/O for all members of the SGI 3000 family. One I-brick comes standard in all entry-level system configurations. The SGI 3000 series supports partitioning to improve system resiliency while maximizing access to your data. A number of these 4U rack-mount enclosures can be added to a system configuration to provide multiple base I/O features and multiple copies of the IRIX(R) operating system. Implementing a configuration that consists of several I-bricks assists customers interested in a very flexible, resilient server with the highest level of availability.

Features:

- One 18GB system disk, optional 2nd drive bay for OS mirroring
- CD-ROM
- 5 PCI expansion slots
- Two Xtown2 1.2GB/second ports (to connect to C-brick)
- Two external USB ports
- One 10/100Base-T Ethernet port
- One external IEEE 1394 port

P-brick

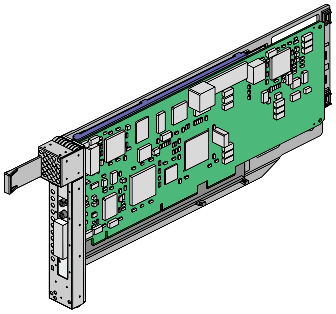


The P-brick is a PCI-based I/O expansion subsystem that is available for customers wanting server configurations with PCI I/O beyond the base I/O offered in the I-brick. Each P-brick is 4U high, comes standard in all SGI 3800 configurations, and is optional on the SGI 3200 and SGI 3400 systems. The P-brick does not merely add PCI slots to the system, it adds incremental I/O to the system. Every P-brick contains three Xbridge chips, each supporting two separate PCI buses, for a total of six PCI buses. Aggregate incremental I/O that is provided by a single P-brick is more than 3GB/second. Two I/O (II) interface connectors in the rear of the unit allow the unit to be dual-hosted to two separate C-bricks, increasing the availability of the I/O in case of a failure.

Features:

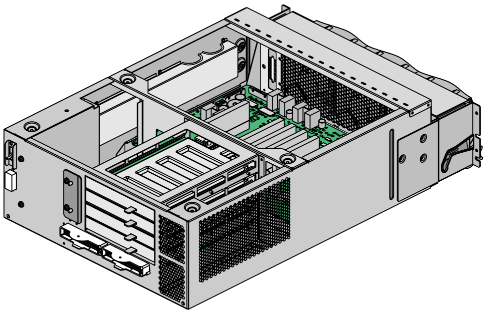
- Provides 12 hot-plug PCI expansion slots
- Provides six 64-bit/66- MHz PCI busses
- Comes with 12 PCI carriers for hot-plug PCI
- Supports 3.3V or Universal PCI cards

PCI Card Carrier



The PCI card carrier is a custom-designed carrier that provides the ability to hot-plug PCI cards in both an I-brick and a P-brick. This feature allows you to add or remove PCI adapters while the system is still in operation, so your ability to change I/O requirements is flexible and dynamic while not affecting application availability. Each carrier supports both full-length or half-length PCI adapters.

X-brick

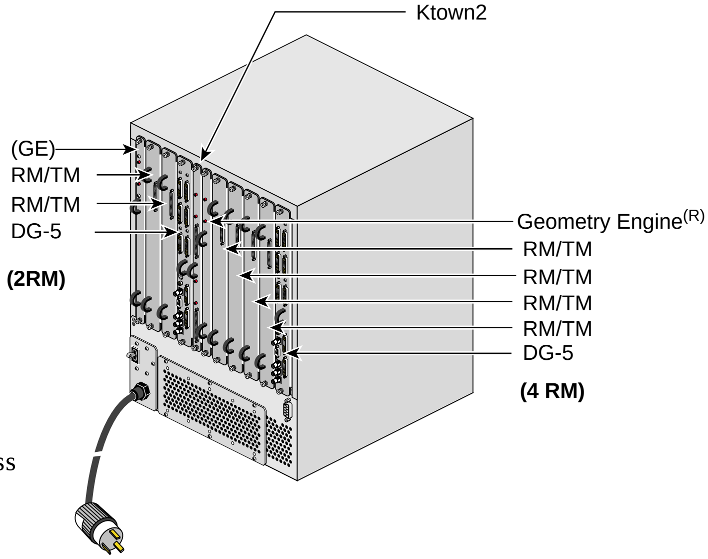


The X-brick is an optional I/O expansion subsystem that supports the XIO interface. Designed to support high-bandwidth applications beyond what the PCI bus can offer, the X-brick is 4U high and supports a number of SGI XIO adapter cards. Each X-brick is powered by a single Xtown2 crossbar, whose aggregate bandwidth is 2.4GB/second, shared between all four XIO slots. Two I/O (II) interface connectors in the rear of the unit allow the unit to be dual-hosted to two separate C-bricks, increasing the availability of the I/O in case of a failure.

Features:

- Provides four expansion slots for standard XIO cards
- Provides two Xtown2 ports
  (1.2 GB/second in each direction)
- Supports many existing XIO cards

G-brick

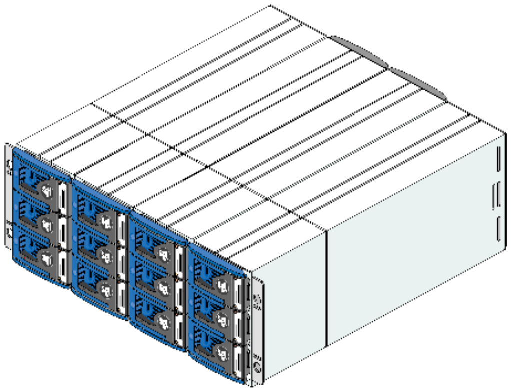


The G-brick is an optional expansion module that houses advanced SGI graphics subsystems, referred to as graphics pipes. The G-brick is optional on all SGI Origin 3000 servers and is part of the standard base configurations of the SGI Onyx 3000 series of systems. The G-brick supports one or two graphics pipes, either InfiniteReality2TM or InfiniteReality3 versions, and has a maximum configuration of one 4RM pipe and one 2RM pipeline. Customers with existingOnyx2(R) InfiniteReality(R) class graphics pipelines can simply migrate to the G-brick by ordering an upgrade kit. The G-brick measures 18Uand requires both one C-brick and one I/O brick for each pipeline. In the SGI Onyx 3400 and 3800 systems, up to two G-bricks can be configured in a graphics rack, each with two pipes, supporting solutions up to 8 pipes and 16 pipes per system, respectively.

The G-brick with graphics pipeline isn't just an adapter that plugs into an I/O slot in the machine, but a tightly integrated graphics subsystem that has a special high-bandwidth I/O interface to deliver the extreme levels of performance necessary to generate sophisticated visualization solutions.

D-brick

The D-brick provides JBOD storage capacity for the SGI 3000 family. The D-brick can be integrated into single-rack SGI 3200 and SGI 3400 systems or into the I/O rack of SGI 3800 systems.



Features:

- Maximum 12 dual-ported Fibre Channel disk drives
- Hot-pluggable drive carriers
- Mounts in standard 19-inch rack, 4U high
- Single-phase input power is switchable 200 to 230 VAC
- Industry-standard SCA-2 interface connectors
- Two hot-pluggable 400 W power supplies and fans that offer full redundancy
- 18.2GB, 36.4GB, and 72.8GB drive capacities, all 10,000 RPM

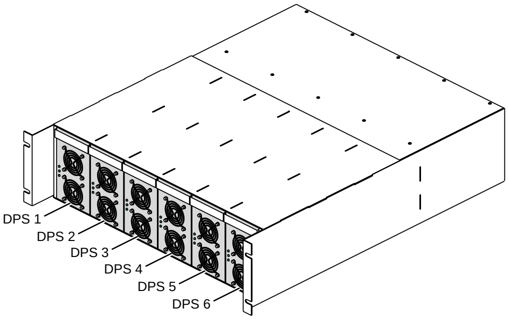
PCI Cards

Marketing Code	Description
PCI-SCSI-LVD-2P	Ultra2 SCSI (80MB / sec) Single-ended 2 port card
PCI-FC-1POPT-A	Fiber Channel 1-Port card with Fiber optic cable
PCI-FC-1PCOP-A	Fibre Channel 1-Port card with copper cable
PCI-GIGENET-C	Gigabit Ethernet (100 / 1000) copper
PCI-GIGENET-OR	Gigabit Ethernet (100 / 1000) fiber optic
PCI-AUD-D1000	Digital Audio
PCI-ATMOC3-1P	ATM OC3 card
PCI-ATMOC12-1P	ATM OC12 card

XIO Cards

Marketing Code	Description
XT-HIPPI-800-SER	Single Port Serial HIPPI card
XT-DIVO	Digital Video card
XT-DIVO-DVC	Digital Video card, supports DVC
XT-GSN-C-1XIO	GSN Adapter single XIO slot, copper
XT-GSN-C-2XIO	GSN Adapter dual XIO slot, copper
XT-HD	High Definition video card
XT-FDDI-D	FDDI Dual Attach single port card
XT-VME-6U	VME card, supports 6U
XT-VME-9U	VME card, supports 9U
XT-ATM-DC3C-4P	Four-port ATM OC3 card

Power Bay



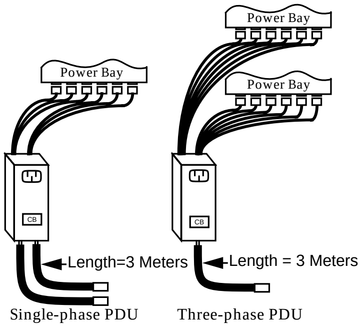
The power bay delivers power to each of the bricks that are used to configure a system. Each power bay is 3U high and houses up to six hot-swap distributed power supplies (DPS). System configurations come standard with a predetermined number of DPSs that provide N+1 redundancy. Smaller configurations come standard with one power bay, while larger or more complex configurations will have two.

Features:

- Provides N+1 power
- Holds a maximum of 6 hot-swap DPSs
- Each DPS is rated at 950 W
- Has eight 48 VDC output connections

Power Distribution Unit (PDU)

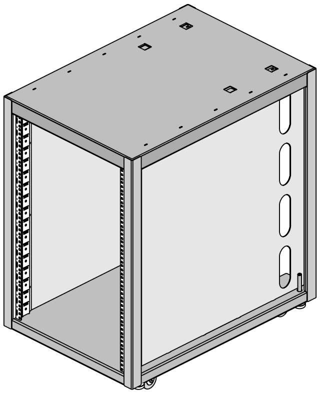
Both single-phase and three-phase power distribution are available.



Single-phase PDU Three-phase PDU

- HU-SN-PDU-1P-NAJ single phase for domestic, Mexico, Canada, and Japan
- HU-SN-PDU-1P-INTL single phase for Europe and elsewhere
- HU-SN-PDU-3P-NAJ three phase for domestic, Mexico, Canada and Japan
- HU-SN-PDU-3P-INTL three phase for Europe and elsewhere

Short Rack

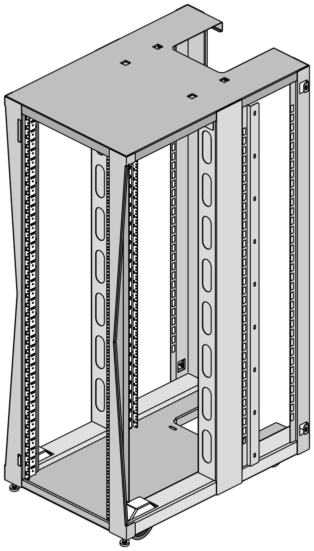


The short rack can be configured multiple ways by combining any of the standard 19-inch rack-mounted bricks up to a total of 17U. The short rack was designed for customers with space limitations or for solutions that are satisfied by scalability to 8 processors. A minimum server configuration would require one power bay, one C-brick, and one I-brick.

Features:

- 17U of configurable space
- 19-inch EIA standard mounting rails
- Mounted on casters
- 36 in. high x 25.5 in. wide x 36 in. deep
- One 220 VAC 10 A power distribution strip

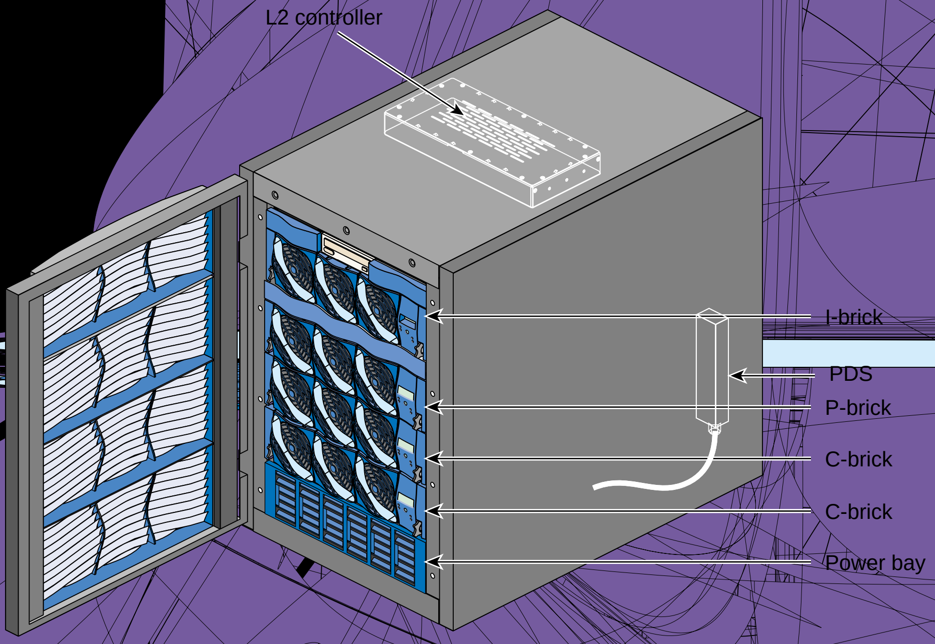
Tall Rack



The 19-inch industry-standard tall rack is the base rack for all highly scalable server configurations. It provides maximum flexibility, allowing customers to design a server that addresses a specific application. An endless combination of bricks can be used to configure an ideal system. Processors, I/O bandwidth, and storage can be added independently to provide a level of granularity and flexibility not offered by any other system architecture.

Features:

- 39 U of configurable space
- 19-inch EIA standard mounting rails
- Mounted on casters
- 74 in. high x 30 in. wide x 50 in. deep
- Can use any of the standard rack PDUs


L2 controller
I-brick
PDS
P-brick
C-brick
C-brick
Power bay

SGI Origin 3200 Technical Specifications

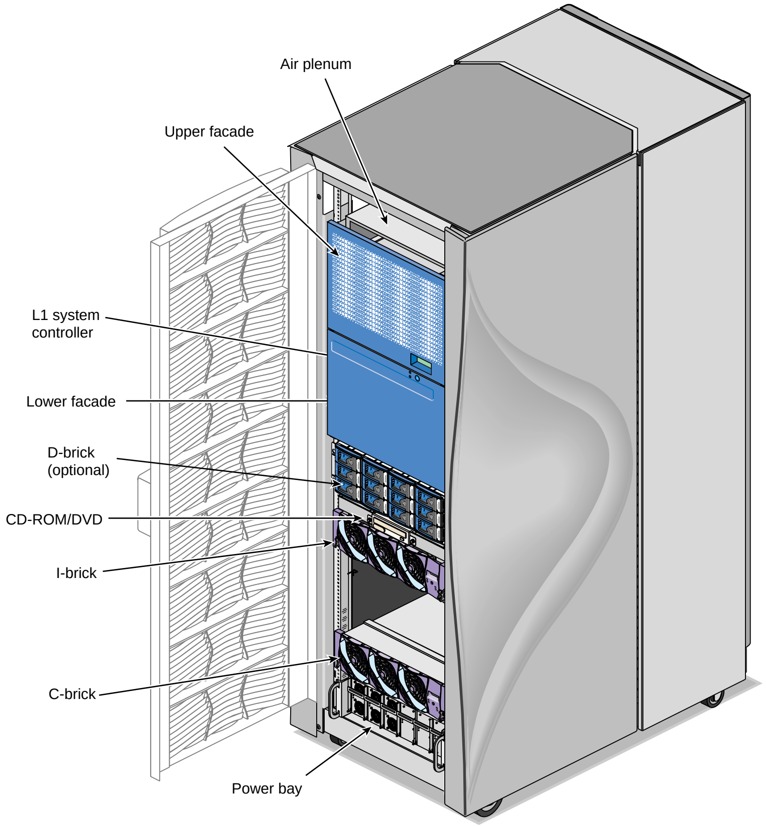
A single 17U deskside rack configuration with no additional racks for processors or I/O:

Description	Minimum	Maximum
Memory size	512MB: (one C-brick with one 512MB bank installed)	16GB: (two C-bricks, each with four 2GB banks installed)
Processors	One C-brick (2 processors)	Two C-bricks (8 processors)
Input/output	One I-brick	One I-brick and one optional I/O brick
R-bricks	None	None
Power bays	One power bay	One power bay

SGI Origin 3200 pre-configured bundles:

Standard Bundles	Processor Base Configurations	Processor Upgrade Possibilities
2 processors	One 2P C-brick	Add (1) 2P PIMM or Add (1) 4P C-brick
4 processors	One 4P C-brick	Add (1) 4P C-brick
4 processors	Two 2P C-bricks	Add (1-2) 2P PIMMS
8 processors	Two 4P C-bricks	N/A

SGI Onyx 3200 System



Although offering the same easy-to-order four bundles as SGI Origin 3200, the SGI Onyx 3200 bundles are all based on the taller 39U rack. Due to the size of the G-brick (necessary for graphics pipes), customers cannot upgrade SGI Origin 3200 servers with a graphics subsystem. All SGI Onyx 3200 system configurations are single-rack solutions, a benefit to customers interested in minimizing space. The base configuration comes with a single G-brick, to which one or two graphics pipes can be added. Additional G-bricks cannot be added. The G-brick also supports the graphics pipes from legacy Onyx2 systems, for customers wanting to migrate their existing graphics to a newer CPU architecture. As with other models of the SGI Onyx 3000 series, InfiniteReality2 or InfiniteReality3 graphics pipes are supported.

SGI Onyx 3200 Technical Specifications

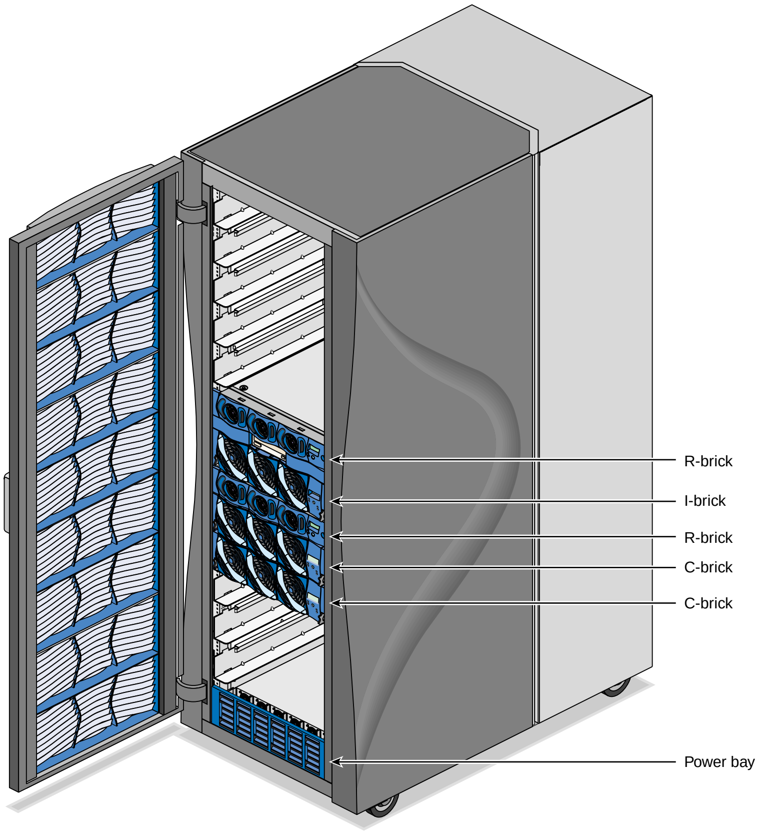
A single 39U rack configuration with no additional racks for processors or I/O:

Description	Minimum	Maximum
Memory size	512MB: (one C-brick with one 512MB bank installed)	16GB: (two C-bricks, each with four 2GB banks installed)
Processors	One C-brick (2 processors)	Two C-bricks (8 processors)
Graphics module	One G-brick (pipe not included)	One G-brick
Graphics pipe	One InfiniteReality2 or IR3 pipe, one RM, one DG, one monitor	Two graphics pipes
Graphics options	Keyboard, mouse, audio card, USB extender	Two sets of options, one for each graphics pipe
Input/output	One I-brick	One I-brick and one optional I/O brick
R-bricks	None	None
Power bays	One power bay	One power bay

SGI Onyx 3200 pre-configured bundles:

Standard Bundles	Processor Base Configurations	Processor Upgrade Possibilities
2 processors	One 2P C-brick	Add (1) 2P PIMM or (1) 4P - C-brick
4 Processors	One 4P C-brick	Add (1) 4P C brick
4 Processors	Two 2P C-bricks	Add (1-2) 2P PIMMS
8 Processors	Two 4P C-bricks	N/A

SGI Origin 3400 Server



The SGI Origin 3400 server comes in one of four easy-to-order preconfigured bundles. Based on the tall rack, SGI Origin 3400 integrates C-bricks, R-bricks, and any of the three I/O bricks into a single rack. For maximum I/O and processor configurations, an optional second I/O-only rack can be added. Two R-bricks with 6-port routers come standard in all configurations, which allow simple upgrades to a maximum of eight C-bricks. Four C-bricks are connected to each R-brick, and the two R-bricks are tightly linked together with two NUMAlink cables for maximum bandwidth.

SGI Origin 3400 Technical Specifications

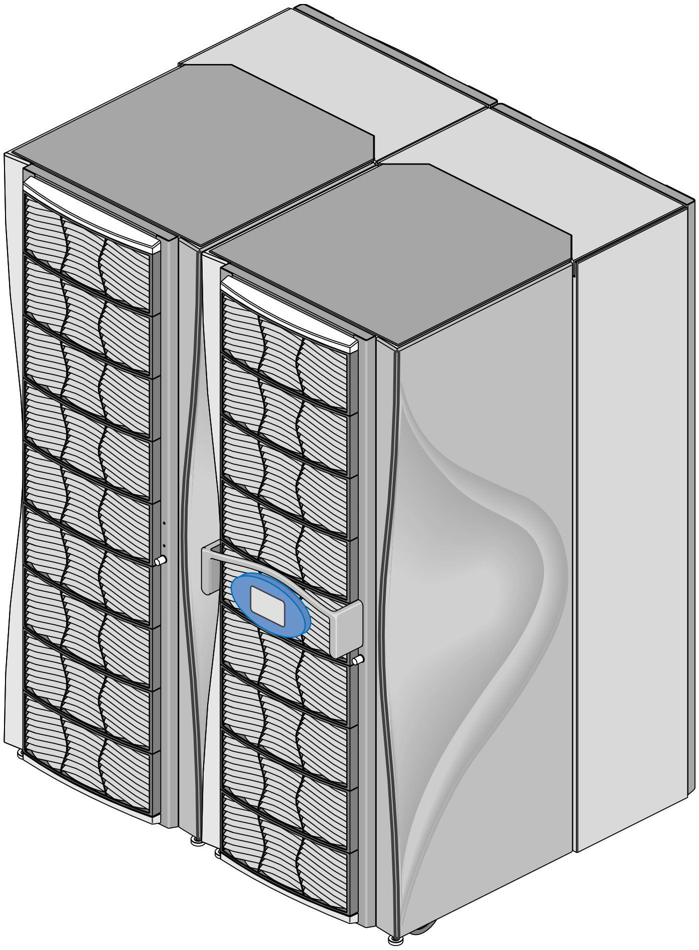
A single 39U rack configuration with additional racks for I/O and disks as required:

Description	**Minimum**	**Maximum**
Memory size	512MB (one C-brick with one 512MB bank installed)	64GB (eight C-bricks, each with four 2GB banks installed)
Processors	4 processors (one C-brick)	32 processors (eight C-bricks)
Input/output	One I-brick	One I-brick, seven optional I/O bricks
R-bricks	Two 6-port R-bricks	Two 6-port R-bricks
Power bays	One power bay	Two power bays

C-brick processor configuration options:

Standard Bundles	**Processor Base Configurations**	**Processor Upgrade Possibilities**
4 processors	One 4P C-brick and 2 R-bricks	Add 1-7 4P C-bricks
8 processors	Two 4P C-bricks and 2 R-bricks	Add 1-6 4P C-bricks
16 processors	Four 4P C-bricks and 2 R-bricks	Add 1-4 4P C-bricks
32 processors	Eight 4P C-bricks and 2 R-bricks	N/A

SGI Onyx 3400 System



The SGI Onyx 3400 system comes in one of four easy-to-order preconfigured bundles. Based on the tall rack, SGI Onyx 3400 integrates C-bricks, R-bricks, and any of the three I/O bricks into the first rack, and a minimum of one G-brick in a second graphics rack. For maximum I/O and processor configurations, an optional third I/O-only rack can be added. Two R-bricks with 6-port routers come standard in all configuration, which allow simple upgrades to a maximum of eight C-bricks, all in the same first (CPU) rack. SGI Onyx 3400 supports up to 4 G-bricks and 8 graphics pipes, based on InfiniteReality2 or InfiniteReality3 graphics.

SGI Onyx 3400 Technical Specifications

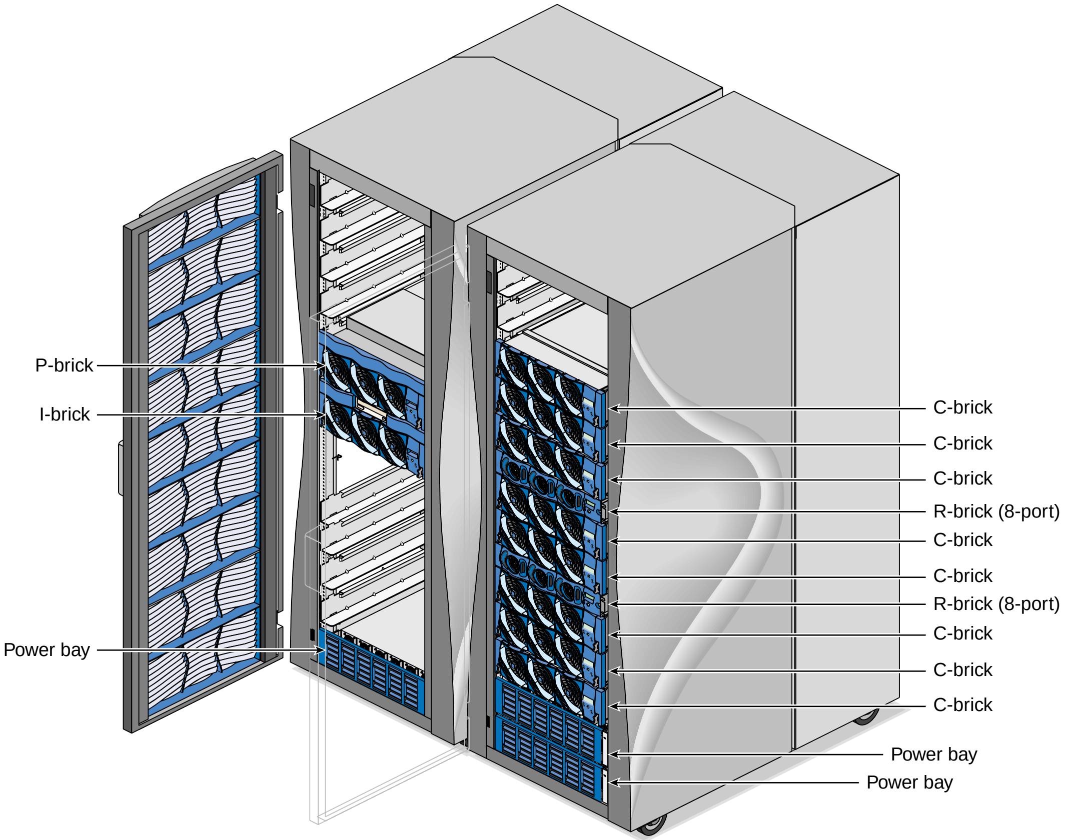
A two 39-U rack configuration with additional racks for I/O and disks as required:

Description	Minimum	Maximum
Memory size	512MB (one C-brick with one 512MB bank installed)	64GB (eight C-bricks, each with four 2GB banks installed)
Processors	4 processors (one C-brick)	32 processors (eight C-bricks)
Graphics Module	One G-brick (pipe not included)	4 G-bricks
Graphics pipe	One InfiniteReality2 or IR3 pipe, one RM, one DG, one monitor	8 graphics pipes
Graphics options	Keyboard, mouse, audio card, USB extender	8 sets of options, one for each graphics pipe
Input/output	One I-brick	One I-brick, seven optional I/O bricks
R-bricks	Two 6-port R-bricks	Two 6-port R-bricks
Power bays	One power bay	Two power bays

C-brick processor configuration options:

Standard Bundles	Processor Base Configurations	Processor Upgrade Possibilities
4 processor	One 4P C-brick and 2 R-bricks	Add 1-7 4P C-bricks
8 processor	Two 4P C-bricks and 2 R-bricks	Add 1-6 4P C-bricks
16 processor	Four 4P C-bricks and 2 R-bricks	Add 1-4 4P C-bricks
32 processor	Eight 4P C-bricks and 2 R-bricks	N/A

SGI Origin 3800 Server



The SGI Origin 3800 server entry configuration includes two R-bricks and four C-bricks in the compute rack and one I-brick and one P-brick in the second (I/O) rack. The two R-bricks have 8-port routers, which are standard in all SGI Origin 3800 configurations. Upgrades can be added in single C-brick increments or by easy-to-order 32-processor full racks. Configurations include more than 128 processors require an additional R-brick called the Metarouter, along with extra NUMAlink cables, to scale the system up to a maximum of 128 C-bricks or 512 processors.

SGI Origin 3800 Technical Specifications

A multiple 39-U rack configuration with additional racks for I/O and disks as required:

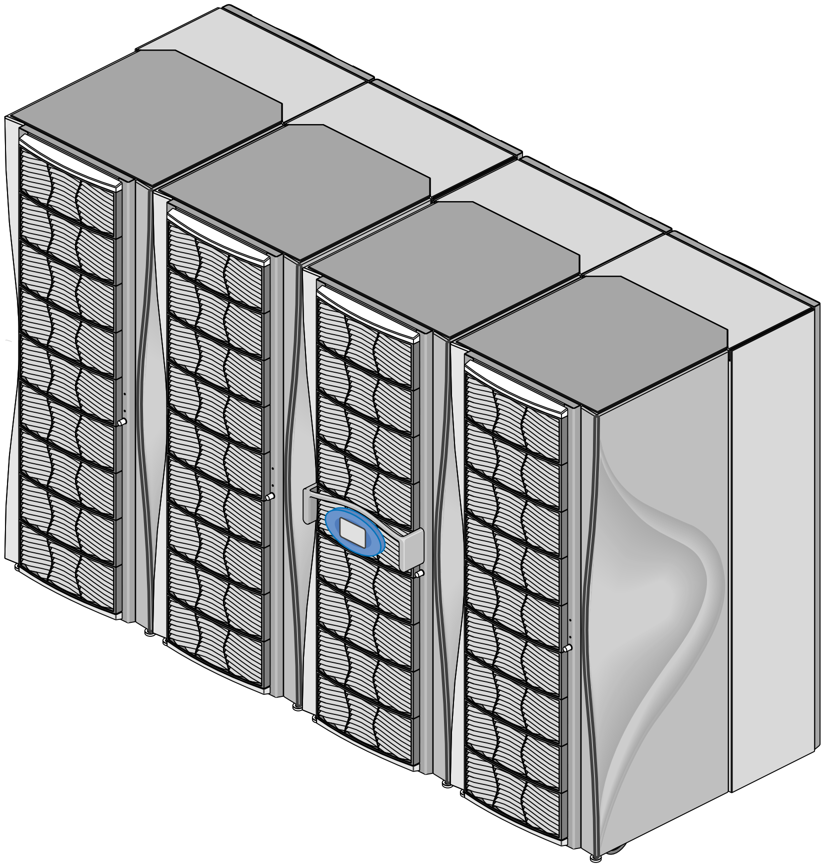
Description	**Minimum**	**Maximum**
39U high racks	One compute rack and one I/O rack	16 Compute racks 8 I/O racks Disk racks as needed
Memory size	2GB (four C-bricks each with one 512MB bank installed)	1TB (128 C-bricks, each with four 2GB banks of memory installed)
Processors	Four C-bricks (16 processors)	128 C-bricks (512 processors) 8 C-bricks maximum per rack
Input/output	One I-brick, one P-brick standard	64 I/O bricks 8 I/O bricks maximum per rack
R-bricks	Two 8-port R-bricks	32 8-port R-bricks, 12 Metarouters
Power bays	One power bay per rack	Two power bays per compute rack One power bay per I/O rack

- Compute racks are configured with only C-bricks and R-bricks.
- I/O racks are configured with only D-bricks and I/O bricks.
- Nine D-bricks per I/O rack is the maximum (no power bays in rack)
- Processors are increased in increments that are based on the size of the system:

  -16 processors to 128 processors in increments of 4, 16, or 32 processors

  –128 processors to 512 processors in increments of only 32 processors
- Configurations that have more than 128 processors require metamemory DIMMs, which include directory memory
- The L2 system controller is mandatory for all compute racks
- The L3 system controller is optional

SGI Onyx 3800



Except for the addition of one or more graphics subsystems, the SGI Onyx 3800 entry configuration is identical to SGI Origin 3800. The base configuration includes two R-bricks and four C-bricks in the compute rack, one I-brick and one P-brick in the second (I/O) rack, and one G-brick housed in a third graphics rack. In addition, for each InfiniteReality2 graphics subsystem, the customer will also receive one keyboard and mouse, one audio card, and one USB extender. SGI Onyx 3800 systems can support up to 16 graphics pipelines. Each graphics rack supports two G-bricks, each with two graphics pipes, for a maximum of four graphics pipes per graphics rack. Like SGI Origin 3800, SGI Onyx 3800 supports up to 512 processors.

SGI Onyx 3800 Technical Specifications

A multiple 39-U rack configuration with additional racks for I/O and disks as required:

Description	Minimum	Maximum
39U high racks	One compute rack, one I/O rack, and one graphics rack	16 compute racks, 8 I/O racks 4 Graphics racks, Disk racks as needed
Memory size	2GB (four C-bricks each with one 512MB bank installed)	1TB (128 C-bricks, each with four 2GB banks of memory installed)
Processors	Four C-bricks (16 processors)	64 C-bricks (512 processors) 8 C-bricks maximum per rack
Graphics module	One G-brick (pipe not included)	8 G-bricks
Graphics pipe	One InfiniteReality2 or 3 pipe, one RM, one DG, one monitor	16 graphics pipes
Graphics options	Keyboard, mouse, audio card, USB extender	16 sets of options, one for each graphics pipe
Input/output	One I-brick, one P-brick standard	128 I/O bricks 8 I/O bricks maximum per rack
R-bricks	Two 8-port R-bricks	32 8-port R-bricks, 12 Metarouters
Power bays	One power bay per rack	Two Power Bays per compute rack One Power Bay per I/O rack

- Compute racks are configured with only C-bricks and R-bricks
- I/O racks are configured with only D-bricks and I/O bricks
- Graphics racks are configured with only G-bricks, 2 per rack
- Nine D-bricks per I/O rack is the maximum (no power bays in rack.
- Processors are increased in increments that are based on the size of the system:

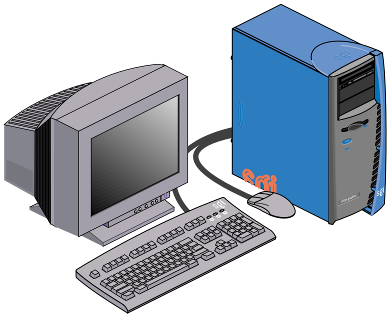
  -16 processors to 128 processor in increments of 4, 16, or 32 processors

  -128 processors to 512 processor in increments of only 32 processors

- Configurations that have more than 128 processors require metamemory DIMMs, which include directory memor.
- The L2 system controller is mandatory for all compute racks
- The L3 system controller is optional

System Management Processor

An optional system-level controller based on a desktop workstation that runs Linux(R) offers you the ability to manage large systems or clusters from a single administrative standpoint. Specialized system diagnostic software performed at this station allows SGI service personnel or customer system operators the ability to run complex diagnostics on the server. The ability to diagnose and predetermine system failures helps the customer avoid unexpected and costly downtime due to parts failure.



Benefits of SGI NUMA 3 Advanced Memory Design

Memory latency is critical to achieve maximum efficiency in a scalable architecture, and significant effort was expended on the SGI NUMA 3 architecture to drive memory latencies extremely low. The non-uniform memory access in SGI NUMA 3 is significantly lower than other NUMA implementations in the industry and also better than many Uniform Memory Access (UMA) architectures. Not only are local and remote access times significantly improved over previous generations of NUMA implementations from SGI, the ratio of remote to local memory is much closer, representing a memory performance very close to what would be expected from an UMA design. The ratio of remote to local memory access in the SGI 3000 family is 2 to 1, whereby the time it takes a processor to access the most remote memory location in the largest configuration (512-processor configuration) is only twice the amount of time it takes to access local memory in the smallest configuration. Memory access times for varying system sizes are listed in the table below for comparison.

SGI 3000 Series Memory Latency

# CPUs	Router Hops Max	Router Hops Average	Local Memory Latency	Worst-Case Remote Latency	Average Latency
16	1	0.75	175	285	257.5
32	2	1.38	175	335	296.3
64	2	1.69	175	335	315.6
128	4	2.47	175	435	356.6
256	5	3.48	175	485	408.3
512	7	4.74	175	585	471.6

Partitioning a System for Resiliency

The SGI 3000 family of systems has an innate ability to deliver high levels of resiliency and availability. The very flexible and modular SGI NUMA 3 architecture is designed with reliability, availability, and serviceability in mind, which customers can use to define a server solution that closely matches their application requirements. With the introduction of SGI NUMA 3, servers from SGI can now be partitioned into separate nodes to create an application environment that emulates a “cluster in a box.”

Partitioning is defined as an ability to take a single distributed shared-memory (DSM) system, any model of the SGI 3000 family for instance, and divide it into a collection of smaller systems. The two primary characteristics of partitioning are:

- The ability to run individual partitions, whereby each partition runs in its own protected memory space with its own operating system kernel and behaves as a distinct, standalone system. Partitions can be booted, powered up or down, and rebooted without affecting the normal operation of the other partitions in the system.
- The partitions are tightly coupled, through the use of the system’s interconnection network (NUMAlink), as a low-latency, high-bandwidth interconnect. A failure that causes a kernel in one partition to panic will not cause a kernel in another partition to crash.

Partitioning can be thought of as a tightly coupled cluster that uses the low-latency and high-bandwidth NUMAlink Interconnect instead of a low-performance networking interface for the interconnect. The higher level of performance that results is directly related to the use of NUMAlink to deliver information between partitions.

Fault isolation is one of the major reasons for partitioning a system. A software or hardware failure in one partition should be isolated from other partitions so that application availability in those other partitions is not affected. To accomplish this, a set of conventions is used to help define partitioning of systems by using IRIX 6.5:

- The minimum configuration for a partition is the combination of one C-brick, one I-brick, and possibly a separate power supply setup. Assuming all C-bricks are fully populated, the minimum partition size would be 4 processors and therefore is the minimum level of hardware isolation.
- Each partition must have the infrastructure to run as a standalone system. This infrastructure includes a system disk and console connection, both located in the I-brick.
- I/O bricks belong to the partition that the attached C-brick belongs to. If an I/O brick is dual-ported to two separate C-bricks, both of these C-bricks must be in the same partition. I/O bricks cannot be shared by two partitions.
- To allow communication to be independent, all intrapartition communication must be through a route that is contained within the partition.
- When the full system is greater than 64 C-bricks (256P), the minimum partition size is four C-bricks (16P, all C-bricks connected to a single R-brick).

IRIX Advanced Cluster Environment (ACE)

As some application environments benefit from a workflow that involves many small shared-memory servers, the ability to manage multiple servers in a simple but effective manner is the key to enhanced productivity. The IRIX Advanced Cluster Environment (ACE) line of products is specifically designed to help system managers improve work efficiency by simplifying the effort to manage many servers, both as individual nodes in a cluster and as multiple partitions in a larger shared-memory server. Used in cluster accounts for years, the tools in ACE for IRIX 6.5 are fully supported for production environments and deliver a single administrative image and single system view of a cluster or a multipartitioned shared-memory system. ACE for IRIX provides many features to effectively manage resources in a production environment:

- Message Passing Toolkit (MPT) provides a method to program scalable computer systems and arrays of workstations and servers. It includes the MPI development environment.
- Job management is supported through the use of Miser, a batch job scheduling facility that balances batch and interactive CPU and memory usage in a single node. The optional Load Sharing Facility (LSF) tool is used to assign effective workload distribution and job scheduling across multiple nodes.
- To manage a set of processes that run on different nodes in a cluster, Array Services provides a way to service conceptually related processes as a single "job."
- To quickly identify and eliminate performance bottlenecks in a cluster, Performance Co-Pilot™ is a performance monitoring tool that supports monitoring across multiple nodes. Customers can chose from a number of add-ons for specialized monitoring and add more collector and monitor agents across the cluster.
- The SGI management processor provides a centralized control point for system operators to manage and monitor multiple nodes or partitions, and logs their activity.
- Software revision control and replication services can be managed with RoboInst, a tool that automates the process of installing operating systems, patches, and applications across multiple nodes or partitions.
- Enlighten DSM is used by system operators for administration management and monitoring of applications, which enables users to immediately begin managing their environment at a workgroup level.
- Clustered XFS™ (CXFS™) is a full-featured clustered file system that enables unparalleled data access across shared disks in a clustered configuration. CXFS is built upon XFS, a 64-bit journaled file system that delivers the industry-leading performance required for big data environments.